

Textual Fingerprinting with texts from Parkin, Bassewitz, and Leander

Christoph Schommer

University of Luxembourg

Dept. of Computer Science - ILIAS Laboratory

6, Rue Richard Coudenhove-Kalergi, 1359 Luxembourg, Luxembourg

Email: christoph.schommer @uni.lu

Conny Uhde

JW Goethe-University Frankfurt am Main

Dept. of Computer Science and Mathematics

Robert-Mayer-Str. 11-15, D-60486 Frankfurt am Main, Germany.

Email: uhde@cs.uni-frankfurt.de

February 15, 2008

Abstract

Current research in author profiling to discover a legal author's fingerprint does not only follow examinations based on statistical parameters only but include more and more dynamic methods that can learn and that react adaptable to the specific behavior of an author. But the question on how to appropriately represent a text is still one of the fundamental tasks, and the problem of which attribute should be used to fingerprint the author's style is still not exactly defined. In this work, we focus on linguistic selection of attributes to fingerprint the style of the authors Parkin, Bassewitz and Leander. We use texts of the genre *Fairy Tale* as it has a clear style and texts of a shorter size with a straightforward story-line and a simple language.

1 What is it about?

The¹. forensic linguistics is concerned with a verification process for the decryption of texts and the analysis through pattern discovery. In this respect, verification means the usage of existing and well-known stylistic attributes to discover an individual (linguistic) fingerprint. However, it is still a controversial discussion, if such

¹This work has been supported by the University of Luxembourg within the project *TRIAS* - Logic of Trust and Reliability for Information Agents in Science

a linguistic fingerprint is a clear indication per se: stylistic tests assume that typical attributes are directly influenceable by the author and that a certain number of attributes still keep constantly, even though the author changes consciously the behavior or the own style [7]. However, following Dixon and Mannion to their evaluations to the texts of Oliver Goldsmith, it may be observed that an appropriate selection of stylistic attributes take a risk: Goldsmith’s style is characterized by an adaptive fluency, where he adapts his own style in a reported speech of the respective actor. To identify Goldsmith’s characteristic style attributes, Dixon and Mannion compared his essays with those of four contemporary writers. They found out that two of the four writers show a suspicious similarity with Goldsmith as they originate from the same Irish area, living in the English exile [14].

Additionally, stylistic attributes that are influenced by the genre may interfere the individual style [3]. This leads to the conclusion that only texts of the same domain are affiliated with each other. Using the texts of the Nijmegen-corpus, Baayen et al. have analyzed the differences of diverse authors of the same genre as well as the texts of authors who represent different genres: they have found out that texts of the same genre are generally more similar than texts of the different genre that are from the same author.

2 Style Discovery

Stylometry refers to the measurement of the style with the aim to fingerprint a text following a certain number of linguistic attributes, to conclude the authorship of a text and/or to order texts following their chronology [18]. The content, the meaning and the correctness of the text is not concerned. The general ambition is to discover those attributes that difference texts sufficiently [17]. Generally, the data is analyzed statistically taking numerical attributes into account but disregarding categorical attributes. Figure of speeches like metaphor and symbols are clearly defined indeed, but are not to be discovered automatically. In [18], Oakes writes that any linguistic occurrence can be taken for the stylometric analysis as the attribute can be expressed by a numerical attribute. However, it must be assured that the attribute is relevant for other genres as well [16]. Another important aspect is the differentiation of linguistic attributes of whether they are consciously controlled by the author or not [13]. Many examinations take explicitly unconscious stylistic attributes as the relevant discrimination criterion as they are a stronger sign of a stylistic fingerprint. However, this includes the existence of stylistic attributes that stay constantly through the whole text and the existence of linguistic attributes that adapt [12]. In this respect, we focus on a differentiation of conscious and unconscious stylistic attributes, well noting that diverse authors differ more in their style than texts of an individual author. Furthermore, texts of an individual author differ more than passages within a text [6]. We therefore conclude that an appropriate consistence of

a continuous usage of conscious and unconscious stylistic attributes must be generally secured. Very generally, linguistic attributes refer either to a statistic frequency or to the differentiation of the vocabulary [20] - under the assumption that authors differ in their vocabulary and that they control their vocabulary rather limited than specific. The vocabulary is then queried by habit, it is performed automatically and therefore constant, appropriate for text classification [10].

Several examinations have shown that a few stylistic attributes are insufficient for the differentiation of authors as they produce *pairs of authors* classifying in the same category. [6] and [19] suggest a wider spectrum of attributes leading to a better success and argue that the selected attributes can be ordered depending on their significance in respect to a classification the genre.

The research on a stylistic analysis for author profiling has been started years ago, when Mendenhall [23] and Mascol [4] examined literary verses of the New Testament by considering aspects like frequency of words the length of sentences. They assumed that authors produce different texts, with different style and features. Many statistical examinations followed, for example to discover text features that may appear constantly. Many attributes have been found and mathematical issues proven, for example the Yules characteristics, Zipf's law and the Hapax Legomena. [15] has shown that a statistical relevance on a low number of textual data can be expressed and computed by a Bayesian statistics, which makes it applicable for a contribution to author profiling. However, to conclude that a text is written by a specific author has not often clear and misclassifications has been done.

Since that time, many other attributes have been examined, for example the number of words of a certain wordclass [1], [9], syntax analysis [2], [21] and word phrases [8], and grammatical failures [11]. Many examinations combine some of these attributes as well as different methods from statistics and machine learning, for example principal component analysis, support vector machines, and cluster analysis [5].

3 About the evaluation environment

In this work, we understand *Author Profiling* as a way to identify authors by a certain number of linguistic (numerical) attributes and to assign texts correctly to them. In this respect, our hypothesis is that - based on the assumption that there exists a potential style identification - an stylistic identification of authors can be done with quality if we can find a sufficient number of expressive attributes, which describe the author's behavior in respect of characteristic and dissimilarity; and that allows an application of machine learning methods for an demonstrative evaluation. Nevertheless, to perform an empirical study in order to discover the author's style is mostly characterized by a linguistic detail, namely the principal use of attributes that are applicable within a computer-based analysis. And an independence of these attributes must be adjusted as well.

3.1 The genre we use

We focus on several texts from the genre *Fairy Tale*. The texts are in german language. Fairy tales have been selected as they are per se an excellent differentiator to other texts; they are distinguished by a clear style and author-independent. Mostly, fairy tales are of shorter size, they are amusing stories with fantastic content without a reference of time. The storyline is straightforward, the language simple.

Although common speech texts are more difficult to differentiate as they do not have to follow a certain style per se, they are contradictory to the texts of a technical texts serving the presentation and the critical discussion with specific contextual aspects. Common speech texts are non-coded and of daily use, easy understandable, but less syntactically defined. Technical texts are often related to science and therefore underlie certain criterions. The understanding of a technical text is highly depending on the style. Nevertheless, authors try to keep their individual style, taking into account the correct use of orthography, syntax and punctuation. The text is often impersonal and written in present tense. We have therefore started our examinations with a comparison between selected texts from *Fairy Tale* and *Common Speech* and *Technical Language*, respectively. Approximately 10 authors per each genre with 3-5 documents have been selected.

3.2 Attribute Selection

In concern of the attributes, we take into account linguistic attributes as much as they significantly contribute to the author’s style, but filter those out that are dependent from another. For the evaluations, we have used more than 30 attributes from statistics or linguistics, for example

- *Number of Words and Number of distinct words*, where punctuation marks are disregarded.
- *Frequency of personal pronoun*. Depending on the genre, the personal pronoun is assessed; for example, the word *I* receives a higher weight in scientific texts than in fairy tales since we may assume that scientific texts follow a more neutral description (passive) or uses the *We* instead.
- *The word with the highest frequency*.
- *Word length in average*.
- *Record length*. We use this attribute although [22] mentions that the record length is not expressive and applicable as a single attribute. The disadvantage is the author’s control and capability to imitate, especially against the punctuation. This makes it less suitable to older texts. [24] agree that the attribute record length is a weak measurement for the author’s style but is useful when focusing on their distributions.

- *Yules characteristic value k* . This value bases on the assumption that the occurrence of a word is random and underlies the distribution of Poisson. The more the words are repeated, the higher k .
- *Hapax Legomena*, the number of words that occur exactly once in the text. This is to measure the author's disposition to use or to avoid synonyms.
- *Sentence Structure*. This attribute describes the author's disposition to prefer main clauses or subordinate clauses. We measure this by the percentage of hypotaxis in the text.
- *Value of the type-token proportion*. Let n the number of tokens (words) in the text and v the number of different tokens, then the type-token proportion r is the fraction between v and n .
- The *Entropy* of the text. The length of each text source is set to a fixed number of words.

Furthermore, we have concerned with *stop words* and calculated (per text) the number of words occurring exactly once, the stop word itself that occurs most frequently, and its frequency and percentage. Using the thesaurus of the University of Leipzig², we additionally calculate the frequency class, which refers to how often a token occurs in comparison to any occurrence.

3.3 Attribute Filtering

A problem in using the selected attributes is that some of them may dependent from others. Furthermore, this depends on the genre very strict, so that attributes of a genre must be preprocessed individually. We therefore have used the statistical method of *Plots* to pairwise visualize the attribute's behavior. For example, the attributes *Entropy*, *Type-Token Ratio*, and *Average word frequency* are dependent whereas *Number of Hapax Legomena* and *Yules characteristics* are not. Figure 1 shows the preprocessed filtering result of texts from the genre *Fairy Tale*, i.e., the pairwise distribution of independent attributes. The presented plot is symmetric.

4 Selected Fingerprinting Results

We have enriched this calculation with diverse statistical methods like principal component analysis or bivariate statistics to visualize and calculate the most interesting and reliable attributes and have applied machine learning in a different way through demographic clustering or a genetic algorithm. Some evaluations have been done

²see Wortschatz - <http://wortschatz.uni-leipzig.de/>

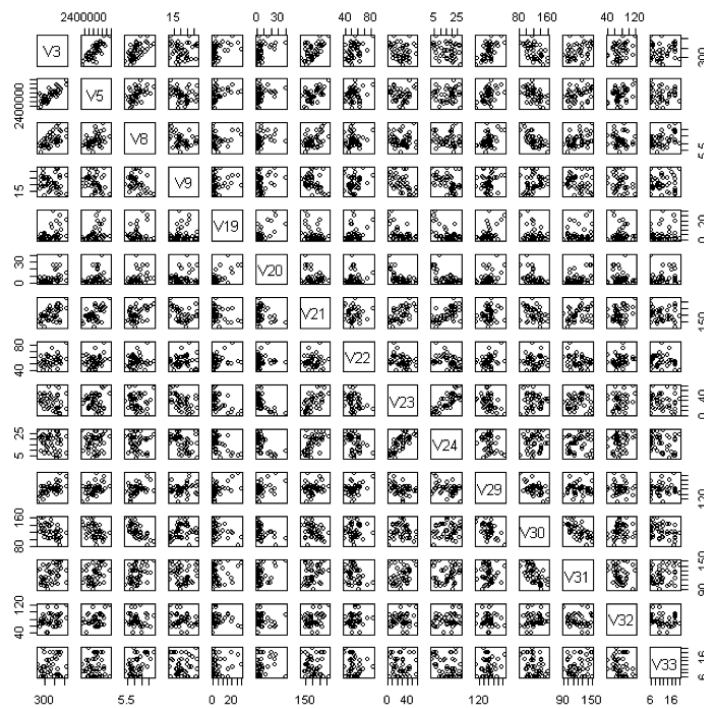


Figure 1: Received list of independent attributes for the genre *Fairy Tale*, plotted in a quadratic and symmetric comparison matrix.

with the *IBM Intelligent Miner V8*, some with the statistical program *R*, and *Clus-Gen*: this a self-programmed genetic simulation to classify the whole text corpus to classes. The idea focus on the assumption that a representative *median vector* for a collection of texts - that come from the same author - dynamically exists; and that the texts of one author are representative enough for all texts of the corresponding author. We have initiated the process of fingerprinting while interpreting all discovered results.

4.1 *Fairy Tale* against *Common Speech* and *Technical Language*

To get an overview of each genre per se and to characterize these texts as well, Figure 2 shows the result of a bivariate statistics against the attributes *Genre*. We observe that 42% of the text set are from *Fairy Tale*, 32% from *Common Speech*, and 26% from *Technical Language*. The variables are ordered following their chi-square values, meaning that the discrepancy between the distribution of the corresponding attribute inside a genre (inner ring for *Genre*, non-colored distribution of the other attributes) to the whole text population (outer ring for *Genre*, colored distribution of the other attributes) represents the significance and therefore position of an at-

tribute in each region: the more different the distribution the more it is positioned to the left.

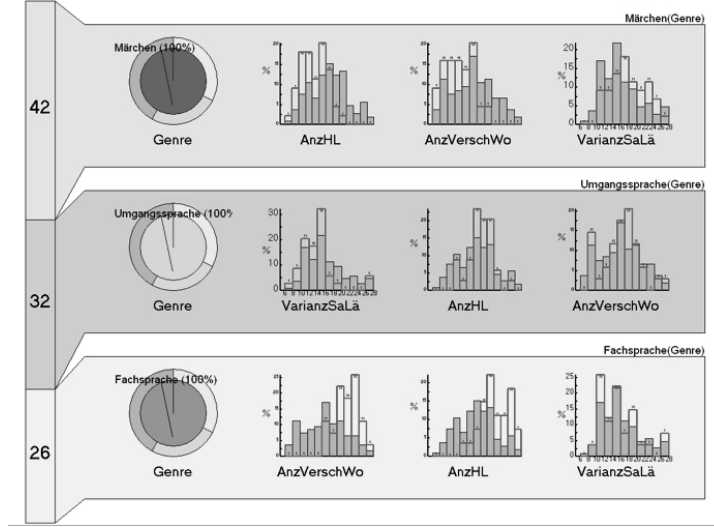


Figure 2: Bivariate Statistics with three variables against the genre (*Fairy Tale* (top), *Common Speech* (middle), and *Technical Language* (bottom)).

In this respect, *AnzHL* represents the most important attribute to *Fairy Tale* (Märchen) and is more significant to its region than in any other genre. Furthermore, the attribute *Number of Hapax Legomena* (*AnzHL*) is characterized by a distribution of high values in *Technical Language*, but is low distributed in *Fairy Tale*. On the other side, the attribute *Number of different tokens* (*AnzVerschWo*) has a distribution with high values in *Technical Language* but with lower values in *Common Language*. Following this, we can assume that in *Fairy Tale* the words occur more seldom only once (*AnzHL*); the number of different words (*AnzVerschWo*) is rather low while having longer sentences (*VarianzSaLä*). Texts from *Technical Language*, however, use a more extended vocabulary (*AnzVerschWo*), where words occur more often only once (*AnzHL*).

With text clustering, we have observed several clusters representing only texts of an individual genre. For *Fairy Tale*, relatively many attributes have a distribution that is surprisingly higher than in all texts, for example the number of adjectives (*AnzAdjektive*), the number of Hapax Legomena (*AnzHL*) and the Yules characteristics (*YulesK*); on the other side, the number of verbs (*AnzVerben*) and the frequency class (*HäufKla*) are quite low distributed. Five authors share this cluster, the most used words are *ich* and *und*. The number of parataxis are relatively lower, the number of hypotaxis relatively higher. We may conclude that the general style is descriptive and figurative, because many adjectives and synonyms are used. It is explainable, since longer sentences exist that are often nested.

Generally, the tests have shown a controversial face. Within the selected texts, the style of only some authors have been constantly, meaning that there exist a set of attributes being equally distributed. This is for the authors *Parkin* and *Bassewitz* as their texts have been selected within a textbook. On the other side, the texts of *Leander* are widespread, although they are also taken from an individual text collection.

4.2 Parkin’s Style

Figure 3 shows a cluster that only bases on texts from the author *Parkin*. We observe a low number of adjectives (*AnzAdjektive*) in all the clustered texts, a high number of parataxis (*proParaTax*), and a high number of *we*’s (*AnzWir*). Parkin’s style is further characterized by a high type-token ratio (*TypeTokenRatio*), a low number of different words (*AnzVerschWo*), and a low number of Hapax Legomena (*AnzHL*). He prefers parataxis, the averaged length of sentences is low as well as the number of different words: he therefore tends to repeat words, favors nouns but not adjectives at all.

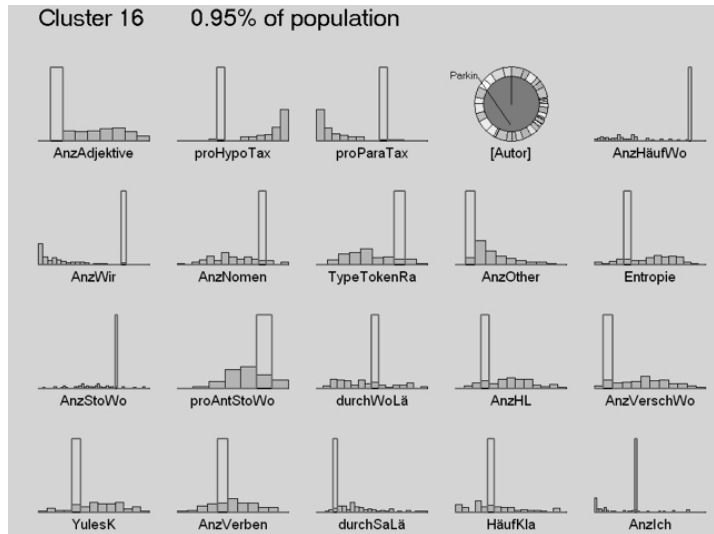


Figure 3: The Parkin-Cluster, Genre *Fairy Tale*.

4.3 Bassewitz’s Style

The Figure 4 shows a snapshot of a demographic clustering result that bases on

- four texts of the authors Bassewitz, taken from *Peterchen’s Mondfahrt*

- six texts of the author Leander, taken from *Träumereien an französischen Kaminen*

within the genre *Fairy Tale*. Bassewitz’s texts are characterized by a high-valued frequency class (*HäufKla*) and a disproportionately high occurrence of verbs. Additionally, the word type of the longest word belongs almost to the same class as well as the word type of the most frequently word, the most frequently stop word, and the number of Hapax Legomena.

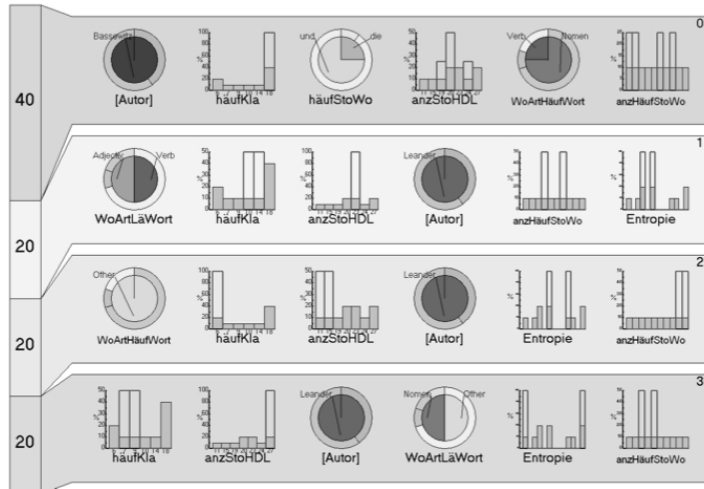


Figure 4: The Bassewitz-Cluster and the Leander-Clusters, Genre *Fairy Tale*.

4.4 Leander’s Style

The texts of Leander are more mixed, not aiming at an uniform style. We observe that Leander’s texts are more different than those from Bassewitz - although both are taken from textbook. Taking the genetic simulation program, we have received classification rates by an adaptive calculation of a new median vector. With this, the classification rates are above 90% for the whole test corpus. Inside the genres, the classification rate for *Fairy tail* has been around 93% but for *Common speech* and *Technical language* nearly 100%.

5 Conclusions

At first glance, it seems not even conceivable that an author’s style or even more, a fingerprint, can be discovered: the number of evaluated texts is quite low and a legal forecast therefore not feasible. But the evaluations prove that through the selection of linguistic attributes, an author can be described within his texts; and even more,

the usage of texts within an author's textbook like Bassewitz's *Peterchen's Mondfahrt* may certainly have its eligibility to characterize his style and to allow to score texts in this microcosm. For the authors *Parkin* and *Bassewitz* this is observed as *Parkin's* style is documented by a couple of attributes sharing an individual distributive behavior, whereas Bassewitz uses - for example - throughout words that occur seldom.

To extend these tests to a larger text corpus, to enrich the given set of attributes by *subjective* linguistic attributes, and to generalize our results to other text corpora will be one of our next responsibilities. In this respect, we understand a *subjective linguistic attribute* as a personal statement of the author himself, like for example *I believe that* or *I certainly agree* to express the personal beliefs and intentions. Furthermore, open questions arise like how representative these results are or how appropriate a scoring engine - that assigns for example a text of *Parkin* correctly - is not tested yet. Our next steps will concern these questions, we also follow up on further examinations. Generally, we strongly believe in this way of style analysis and author recognition, and hope to discover attributes that uniformly relies on our hypothesis.

References

- [1] G. Avneri, S. Argamon, M. Koppel: Routing documents according to their style. Intl. Workshop on Innovative Internet Information Systems, 1998.
- [2] T. F. Baayen, H. v. Halteren: Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, (3):121-130, 1996.
- [3] H. Baayen, H. von Halteren, F. Tweedie: Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, (3):121-130, 1996.
- [4] C. Maskol: Curves of pauline and pseudo-pauline style i+ii. *Unitarian Review* 30:452460, 1988.
- [5] D. Khmelev: Distributed authorship resolution using relative entropy for Markov Chain of letters in texts. 4th Intl Conference on Quantitative Linguistics Association, 2000.
- [6] F. Dimpel: Computergestützte textstatistische Untersuchungen an mittelhochdeutschen Texten. Tübingen: Francke, 2004.
- [7] P. Dixon, D. Mannion: Goldsmith's Periodical Essays: A Statistical Analysis of Eleven Doubtful Cases. *Literary and Linguistic Computing*, 8:1 19, 1993.
- [8] : F. Smadja: Tge missing link. *Journal of the Association for Literary and Linguistic Computing*, 4(3), 1989.
- [9] D. Holmes, R. S. Forsyth: Features finding for text classification. *Literary and Linguistics Computing*, 11(4):163-174, 1996.
- [10] D. L. Hoover: Another Perspective on Vocabulary Richness. *Journal on Computers and the Humanities*, Springer, pp. 151-178, 2004.

- [11] J. Schler, M. Koppel: Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.
- [12] N. M. Laan: Stylometry and Method. The Case of Euripides. Oxford Journals, Literary and Linguistic Computing, pp. 271-278, 1995.
- [13] G. Ledger: Re-Counting Plato: A Computer Analysis of Plato's Style. Clarendon Press. 1990.
- [14] D. Mannion, P. Dixon: Authorship Attribution: the Case of Oliver Goldsmith. Journal of the Royal Statistical Society (Series D): The Statistician, 46:1-18, 1997.
- [15] D. L. Wallace, F. Mosteller: Applied Bayesian and classical inference. Springer, 1984.
- [16] T. Mcenery, M. Oakes: Authorship Identification and Computational Stylometry. Handbook of Natural Language Processing. pp. 545-562, 2000.
- [17] M. A. Queen: Literary Detection. How to prove Authorship and Fraud in Literature and Documents. New York, 1978.
- [18] M. P. Oakes: Statistics for Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press. 1998.
- [19] J. Rudman: The State of Authorship Attribution Studies. Some Problems and Solutions. Kluwer Academic Publishers, 1998.
- [20] E. Stamatatos, N. Fakotakis, G. Kokkinakis: Computer-based Authorship Attribution Without Lexical Measures. Journal on *Computers and the Humanities*, Springer, 35:193-214, 2001.
- [21] G. Kokkinakis, E. Stamatatos, N. Fakotakis: Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4):471-495, 2000.
- [22] M. W. Smith: Recent experience and new developments of methods for the determination of authorship. Association for Literary and Linguistic Computing Bulletin, 11:73-82, 1983.
- [23] T. Mendenhall: The characteristic curves of composition. Science, pp. 214:237249. 1887.
- [24] D. R. Tallentire: An appraisal of methods and models in computational stylistics, with particular reference to author attribution. PhD Thesis, Univesity of Cambridge. 1972.